45054

FUNDAMENTAL AND APPLIED TOXICOLOGY 26, 211–222 (1995)

Dose–Response Assessments for Developmental Toxicity

IV. Benchmark Doses for Fetal Weight Changes[1]

ROBERT J. KAVLOCK,* BRUCE C. ALLEN,† ELAINE M. FAUSTMAN,‡ AND CAROLE A. KIMMEL§

*Developmental Toxicology Division, Health Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina 27711; †K. S. Crump Division, ICF Kaiser, Ruston, Louisiana 71270; ‡Department of Environmental Health, University of Washington, Seattle, Washington, and Affiliate of the Child Development and Mental Retardation Center, Seattle, Washington 98195; and §Human Health Assessment Group, Office of Health and Environmental Assessment, U.S. Environmental Protection Agency, Washington, DC 20460

Received August 22, 1994; accepted December 27, 1994

Dose–Response Assessments for Developmental Toxicity. IV. Benchmark Doses for Fetal Weight Changes. KAVLOCK, R. J., ALLEN, B. C., FAUSTMAN, E. M., AND KIMMEL, C. A. (1995). *Fundam. Appl. Toxicol.* 26, 211–222.

Recently, most attention on the application of benchmark dose (BMD) techniques to toxicology data has focused on quantal measures of response. Before the advantages of the BMD approach can be exploited in the risk assessment process, it is important that continuous measures of response also be modeled appropriately. In this study, we examined a variety of approaches to estimating BMDs for a change in fetal weight following chemical exposure from a total of 85 developmental toxicity experiments. We modeled the change in the mean fetal weight of a litter in response to treatment using a continuous power model, as well as reductions in the weight of individual fetuses within litters (defined as falling below a preset level) using a log–logistic model which incorporates litter size as a covariable and considers intralitter correlations. For the litter-based approach, several methods of defining a benchmark effect (BME) were considered, including a percentage change in mean litter weight, a change in mean litter weight relative to variability in the control group, and a reduction in the mean litter weight to some point on the control group distribution curve. For the fetus-based approach, we examined several BME options on the cumulative frequency distribution of the control fetuses for defining a low weight fetus and calculated several levels of additional risk. BMDs for four litter-based BMEs (a difference of 5% in mean fetal weight, a decrease to the 25th percentile mean weight of control litters, a decrease in the mean weight by 2 standard errors, and a decrease of 0.5 standard deviation units) and two fetus-based BMEs (a 5% added risk of weighing less than the 5th percentile of control weights and a 10% added risk of weighing less than the 10th percentile) showed strong similarities to each other and to statistically derived NOAELs. In addition to providing comparison with the NOAEL as a reference value, these analyses provided confirmation of the advantages of the BMD approach over the NOAEL in terms of the influence of dose spacing and dose selection. Combined with our previous analyses of quantal endpoints of fetal effects, this information provides a firm basis upon which to implement the benchmark dose concept in developmental toxicity risk assessments.

[1] These results were presented in part at the Teratology Society Meeting, Tucson, AZ, June 1993. The views in this paper are those of the authors and do not necessarily reflect the views or policies of the U.S. Environmental Protection Agency.

The benchmark dose (BMD) has been proposed as an advancement over the NOAEL approach for establishing the critical effect in noncancer toxicity studies. This approach, first proposed by Crump (1984), applies a mathematically based dose–response model to the experimental data and estimates a dose corresponding to a predefined level of effect (the benchmark effect or BME), as well as the confidence intervals on that dose. The BMD has been operationally defined as the lower 95% confidence interval on dose for the BME. Advantages of the BMD over the NOAEL include the use of all the experimental dose–response data, less dependency on dose selection and dose spacing, and the rewarding of better experimental designs (more numbers per group, more groups) in the calculation of the confidence limits about the effect level. These advantages have been discussed in a variety of forums and publications (Crump, 1984; Kimmel and Gaylor, 1988; Kimmel, 1990; Gaylor, 1989; Faustman *et al.*, 1994; Allen *et al.*, 1994a,b; California EPA, 1994; Barnes *et al.*, 1995).

We have been critically evaluating the application of the BMD approach to developmental toxicity data using a large compilation of data from standard study designs (Faustman *et al.*, 1994; Allen *et al.*, 1994a,b). These efforts focused on the application of both generic BMD models and models that were specifically developed to analyze quantal endpoints (prenatal death and/or malformations) from developmental toxicity studies. We found that both the generic models and the developmental toxicity-specific models were able to fit the observed dose–response patterns, and that NOAELs determined from the same database were generally similar to

0272-0590/95 $12.00

BMDs estimating the 5 or 10% effect level, depending on the type of data.

The application of BMD methodology in the risk assessment process has been slowed by the limited attention paid to applying the methodology to continuous endpoints of effect such as fetal weight. While the primary focus on BMD methodology has been with quantal endpoints, a few investigators have explored applications to continuous variables. Indeed, in the original description of the approach (Crump, 1984), four different regression models were used to estimate increased risks of liver fat content in carbon tetrachloride-exposed rats, decreased body weights in hexachlorobutadiene-exposed rats, and decreased thymus weights in TCDD-exposed rats. From those three datasets, he concluded that the BMD corresponding to an extra response of 1% was comparable to the NOEL. Gaylor and Slikker (1990) proposed a four-step process to estimate the risk of an adverse effect on neurological function (5-hydroxytryptophan levels in the hippocampus of rats following exposure to methylenedioxymethamphetamine). The first step involved obtaining a mathematical relationship between the mean level of a neurotransmitter in the brain and the administered dose. The second step assumed the distribution of individual measurements of the neurotransmitter around the average values which were estimated from the dose–response relationship. In the third step, an adverse change in the neurotransmitter level was established (in the example they used a concentration equal to three standard deviations below the mean value of the control group), and in the final step the proportion of individuals reaching the adverse effect level was estimated as function of dose. Pease *et al.* (1991) proposed a quantitative risk model for reproductive risk of exposure to dibromochloropropane (DBCP) that incorporated a benchmark dose for sperm counts from exposed rabbits, combined with an approach to low-dose, interspecies extrapolation. They defined the benchmark dose as the lower 95% confidence limit on the dose that produced a mean sperm count in exposed animals that was 10% less than the mean sperm count in control animals. Using data from an epidemiological study, they then postulated that a similar 10% reduction in sperm counts in humans would result in an absolute increase in male infertility of 0.44%. Humans were estimated to be either equal to rabbits in sensitivity to DBCP or up to 60 times more sensitive, and potential increases in infertility in humans from various levels of exposure were calculated and compared with a more standard reference dose. In another approach, Catalano *et al.* (1993) presented a combined analysis of fetal death, weight, and malformation using data from a standard developmental toxicity study. They defined a low weight fetus as one weighing less than three standard deviations below the mean of control animals. To obtain the joint probability of the combined outcomes, they first modeled fetal death as a function of dose and then modeled the outcome of live fetuses in a two-step process (weight as function of dose and malformations as a function of dose and residuals from the weight model). No study to date, however, has examined the influence of various definitions of an adverse effect with respect to BMD estimates for a continuous variable and compared these to the NOAEL.

In this paper, we show that fetal weight, a routinely measured and frequently sensitive indicator of developmental toxicity, can be adequately modeled by both continuous power and log–logistic dose–response models and that several approaches provide BMDs that are similar, on average, to NOAELs. While the limitations of the NOAEL approach make it an imperfect standard by which to judge BMD estimates, such comparisons at least provide a basis with which to compare this technique with historical practices. We also note a fraction of studies in which the BMD calculations differ from the NOAELs by factors exceeding fourfold, largely the result of artifacts of study designs that utilized exceptionally wide dose spacing.

METHODS

The database used in this investigation was a subset of that described by Faustman *et al.* (1994) for the analysis of benchmark approaches to quantal endpoints of developmental toxicity. Datasets from two sources, the National Toxicology Program (NTP) and WIL Laboratories were selected for analysis of the continuous variable fetal weight because study designs used by those two groups recorded individual fetal weights and because the raw data were directly available in electronic format from those sources. These datasets consisted of 173 developmental toxicity studies (96 rat, 56 rabbit, 20 mice, 1 hamster). The highest noneffective dose level for reductions in fetal weight was calculated for each study using a Mantel–Haenszel trend test (Haseman, 1984) and dropping the highest dose in a sequential manner as proposed by Tukey *et al.* (1985). This no-statistical-significance of trend (NOSTASOT) dose was equated with the NOAEL. Previously we had demonstrated that the NOSTASOT technique produced results very similar to NOAELs derived using expert judgment (Faustman *et al.*, 1994). The NOSTASOT dose was less than the highest dose tested, that is there was a dose-related decrease in fetal weight, in 49% (85 of 173) of the studies (Table 1). The preponderance of studies with fetal weight effects was greater in the NTP studies than in the WIL studies, as was the case for other endpoints of developmental toxicity such as malformations and prenatal death (Faustman *et al.*, 1994). This was probably the result of different selection criteria for chemical testing between the two laboratories. The analysis of fetal weight added important new information on the effects of the agents evaluated in the database, as in 31% (26 of 85) of the experiments in which fetal weight was significantly affected, there were no effects on prenatal death or malformations. Conversely, in 36% (32 of 88) of the studies in which there were no effects on fetal weight, significant effects on prenatal death or malformations were noted. Among the three species, it appeared that reductions in fetal weight were less frequently associated with malformations or prenatal death in rabbits than was the case in rats and mice. Thus, in 62% (15 of 24) of the rabbit studies in which there were dose-related effects on malformations or prenatal death, there were no effects on fetal weight, whereas the comparable incidences in rats and mice were 33% (16 of 49) and 6% (1 of 17), respectively. Since there were fewer mouse studies in the database, caution should be applied when deriving firm conclusions from these observations.

To obtain BMD levels for comparison to the NOSTASOTs, we initially applied 18 different definitions of a BME to a subset of 20 NTP studies (7

TABLE 1

Cross-Tabulation of Effects on Prenatal Death and Malformation with Reductions in Fetal Weight in the Database Used to Evaluate Benchmark Dose Approaches for Fetal Weight

Dose-related effect on prenatal death or malformations?	Dose-related decrease in fetal weight? Yes	No	Total
Rats			
Yes	33	16	49
No	21	26	47
Total	54	42	96
Rabbits			
Yes	9	15	24
No	2	30	32
Total	11	45	56
Mice			
Yes	16	1	17
No	3	0	3
Total	19	1	20
All species[a]			
Yes	59	32	91
No	26	56	82
Total	85	88	173[a]

[a] Includes one hamster study (positive for both death/malformation and decreased fetal weight) in addition to those in the rat, rabbit, and mouse.

mouse, 9 rat, and 4 rabbit studies involving a total of 10 test agents). These 18 definitions of a BME can be grouped into 2 main categories: (1) those pertaining to the difference in mean fetal weights between treated and control litters based on some defined magnitude of response (litter-based approaches), and (2) those pertaining to the incidence of low weight fetuses within individual litters (fetus-based approaches). These BMEs thus encompassed a variety of changes in fetal weight with which to contrast BMDs and NOSTASOTs.

Litter-based approaches. For analysis of mean litter weights, a continuous power model was used to identify BMDs from BMEs defined as (1) a reduction in the average litter weight by a set percentage, (2) a reduction in the average litter weight to that of a set percentile of the control distribution; and (3) a reduction in the average litter weight by multiples of the control standard error or standard deviation. The BMEs for defining a litter with a reduced mean fetal weight are presented graphically in Fig. 1 and described in detail below.

The continuous power model can be expressed as

$$m(d) = \alpha + \beta d^{\gamma},$$

where $m(d)$ is the average mean litter weight for dose d, and α, β, and γ are parameters estimated by maximum likelihood methods. Normally distributed average litter weights (with dose-group-specific variances) were assumed for the maximum likelihood calculations. The BMD is defined as the 95% lower confidence limit on the dose estimated to produce the desired level of response where the computation of confidence limits was based on the asymptotic distribution of the likelihood ratio statistic (the likelihood-based limits—Cox and Oakes, 1984; Crump and Howe, 1985). The continuous power model was used previously to calculate BMDs (the CBMDs) for the proportion of implants or fetuses per litter showing prenatal death or malformations (Allen *et al.*, 1994a).

The BME for the percentage reduction in mean litter weight,

$$(m(0) - m(d))/m(0),$$

was set at 5 and 10%.

For approaches based on the absolute change from control,

$$m(0) - m(d),$$

the 5th, 10th, and 25th percentiles of the control weight distribution were used as differences for defining the BME.

For the change relative to variation in the control group,

$$(m(0) - m(d))/\sigma(0),$$

relative differences of 0.05, 0.1, 1, 2, and $2/\sqrt{n}$ were examined with respect to defining the BME. In the last definition, $2/\sqrt{n}$, n is the sample size of the control group; the resulting benchmark is equivalent to a reduction in the average litter weight of two standard errors of the control mean. This approach will be referred to in the text as the "two standard error" difference.

Fetus-based approaches. Modeling of the individual fetal weight within litters was done as a quantal variable using an extension of the log-logistic model (Kupper *et al.*, 1986) that considered litter size and within-litter correlations. This model was applied previously to this database to calculate the fetus-based BMDs (the LBMDs) for prenatal death and malformations (Allen *et al.*, 1994b). The model is expressed as

$$P(d, s) = \alpha + \theta_1 s + [1 - \alpha - \theta_1 s]/[1 + \exp\{\beta + \theta_2 s - \gamma \log(d)\}],$$

where $P(d, s)$ is the probability of a low weight fetus at dose d and litter size s, and parameters α, β, γ, θ_1, and θ_2 are estimated by methods of maximum likelihood. A beta-binomial model (with dose-group-specific correlation coefficients) was assumed for the likelihood calculations. Low weight fetuses were defined as weighing less than the 5th or the 10th percentile of the concurrent control group distribution. Added risk levels of between 1 and 25% were chosen as the BMEs.

BMD estimates for the initial 18 definitions of the BME were compared to one another and to the corresponding NOSTASOTs through the use of ratios and their descriptive statistics (mean, standard deviation, minimum, maximum). Additional descriptive statistical tools (median, frequency distributions, and Spearman rank correlations) were utilized in evaluation of the six BMEs applied to the larger database. These analyses aided in comparison of the BMDs to the NOSTASOTs, and identification of particular studies that yielded results divergent from the main body of the data sets.

RESULTS

Preliminary evaluation of BMEs in subset of database. Lacking a clear definition of a biologically significant effect on fetal weight, we began exploring potential BMEs by selecting several basic approaches and exploring several levels of effect for each approach. For example, one approach was based on the change in mean litter response to some percentile of the control mean litter weight distribution. The mean

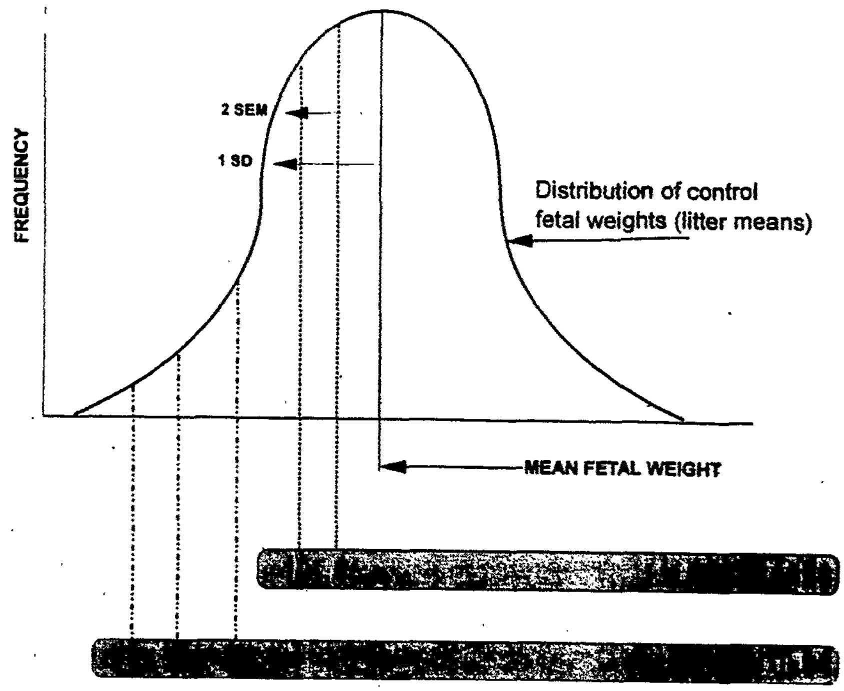

FIG. 1. Graphical representation of the various BMEs used in determination of the litter-based BMDs. The hypothetical normal distribution of control mean litter weights is depicted, along with indications of the locations on the distribution curve for reductions in mean litter weight of 5 and 10%, reductions in mean litter weight equal to the 5th, 10th, and 25th percentiles, as well as reductions equal to the magnitude of the standard deviation and two standard errors of the mean. The positions of the various BMEs relative to each other have been exaggerated for illustration purposes.

fetal weights corresponding to the 5th, 10th, and 25th percentiles in the NTP datasets were 0.80, 0.82, and 0.84 g for the mouse; 30, 33, and 38 g for the rabbit; and 2.8, 2.9, and 3.1 g for the rat, respectively (see Figs. 2A–2C). Another approach was based on defining the individual fetal weights as being reduced in growth relative to the control fetal weight distribution. Fetal weights for the percentiles that best matched the NOSTASOTs (the 5th and 10th percentiles) in the NTP datasets were 0.78 and 0.81 g for the mouse, 26 and 30 g for the rabbit, and 2.6 and 2.9 g for the rat, respectively (Figs. 2D–2F). Both the litter- and fetal-based percentile BMEs lie on the lower end of the cumulative distribution curves, but in the latter instance, they were much closer to the lower inflection point.

For the 18 definitions of the BME shown in Table 2, the mean ratio of the BMDs to corresponding NOSTASOTs ranged from a low of 0.1 for the litter-based approach of a change of 0.05 standard deviations from the control mean to a high of 3.5 for a change in the mean of 2 standard deviation units. Many of the fetus-based BME definitions provided BMD/NOSTASOT ratios near unity, although those based on fetuses weighing less than the first percentile were larger than those based upon higher percentiles. This might reflect greater uncertainty in observing responses at very low ends of the weight distribution. Larger differences in the BMD/NOSTASOT ratios were observed within the various litter-based BME definitions, and selection of those with values near unity was straightforward. The exceptions to this were the BMEs based upon change relative to the control standard deviation, where our options were either less than or greater than the corresponding NOSTASOT. Thus, a value of 0.5 SD was used for subsequent evaluations.

Application of six BMEs to larger database. From these 18 possibilities, we selected 6 that provided the closest similarity to the NOSTASOT for application to the full complement of studies from the NTP and WIL Laboratories. BMDs derived from these six BMEs (the shaded options in Table 2) included: (1) the dose reducing the average mean litter weight by 5%, (2) the dose yielding an average mean litter weight equivalent to that of the 25th percentile of the control mean distribution; (3) the dose yielding an average decrease in mean litter weight equal to twice the standard error of the control group mean; (4) the dose yielding an average mean litter weight 0.5 standard deviation units below the control group mean (this BME fell between the 0.1 and 1 standard deviation unit BME definitions used in the preliminary study and was believed to more closely resemble the NOSTASOT; see Table 2); (5) the dose yielding a 5% increase in the expected proportion of fetuses weighing less than the 5th percentile of the control weight distribution; and (6) the dose yielding a 10% increase in the expected proportion of fetuses weighing less than the 10th percentile

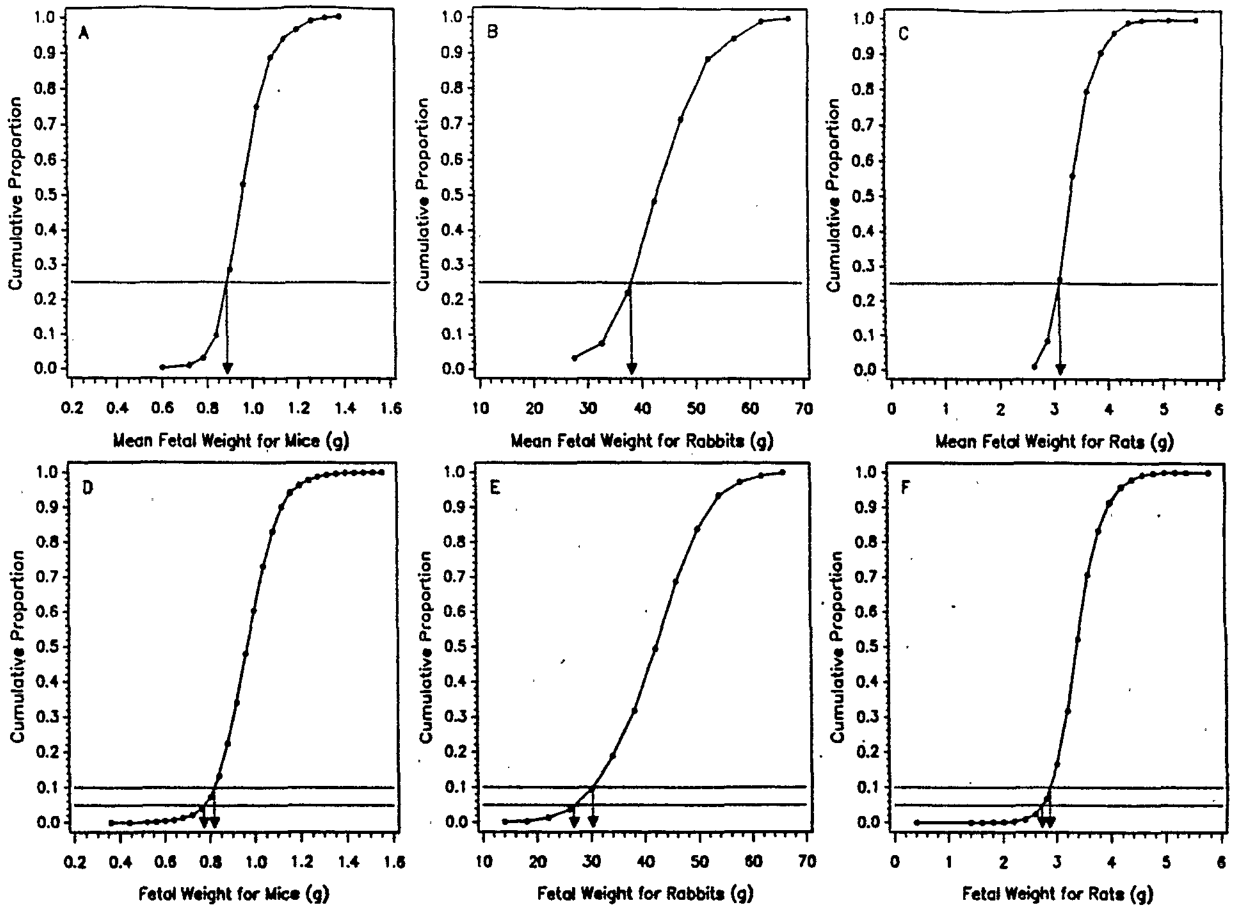

FIG. 2. Individual litter (A, B, and C) and fetal (D, E, and F) weight distributions for control mice, rats, and rabbits in the NTP datasets. In A–C, the horizontal dotted line is drawn at the 25th percentile of the mean fetal weight in control litters (the percentile with the closest similarity with the NOSTASOT) and the vertical line indicates the associated weight. Similar dotted lines are drawn in D–F at the 5th and 10th percentiles of individual fetal weights (see Results).

of the control weight distribution. The latter two benchmarks were estimated using the log–logistic model, while the first four were estimated using the continuous power model.

In 9 of the 85 datasets that had significant dose-related effects on fetal weight, the NOSTASOT test identified a significant reduction in fetal weight at the lowest non-zero dose; therefore, direct comparisons between the BMDs and NOSTASOTs were possible for only the remaining 76 datasets. In a few other datasets, the log–logistic model failed to converge on a risk estimate, and comparisons between BMDs and NOSTASOTs were further reduced for the fetus-based BMEs. This occurred in two datasets for the BME based upon the 5th percentile, and in three different datasets for that based upon the 10th percentile.

Summary statistics of ratios of the BMDs/NOSTASOTs for each approach used to calculate BMDs for fetal weight are provided in Fig. 3, with individual frequency distributions displayed in Fig. 4. In general, all approaches yielded BMD to NOSTASOT ratios with median values near unity (range 0.90 to 1.27). Mean values of the ratios of the BMDs to NOSTASOTs were greater than the median values (range 1.34 to 2.10), indicating that some benchmark values were considerably greater than the corresponding NOSTASOT (i.e., the distribution of the ratios were skewed to values above unity—see Figs. 3 and 4).

Using ratios between 0.5 and 2 of the BMD to the NOSTASOT as a measure of similarity between the two values, concordance ranged from a high of 85% (65/76) for the two standard error approach to a low of 67% (51/76) for the 5% reduction in mean litter weight. The number of datasets in which the ratio of the BMD and NOSTASOT exceeded a factor of 4 ranged from a low of 3 using the 2 standard error BME to 9 using a 5% decrease in mean litter weight as the BME. The largest individual difference was a factor of 18 using the 25th percentile as the BME (Fig. 3, Table 3). In no instance was a BMD value less than one-fourth the NOSTASOT.

TABLE 2

Summary Results of Various Pilot BME (Benchmark Effect Levels) Approaches Used to Calculate BMDs (Benchmark Doses) for Fetal Weight from a Survey of 20 NTP Studies[a]

	Litter-based approaches										Fetus-based approaches							
	% Decrease in litter mean		Dose mean equal to control percentile			Change relative to control standard error	Change relative to control standard deviation[b]				Incidence of low weight fetuses (as percentile-added risk)							
	5	10	5	10	**25**	**2 SEM**	0.05 SD	0.1 SD	1 SD	2 SD	1–5	1–10	**5–5**	5–10	10–5	**10–10**	25–5	25–10
Mean	1.3[c]	2.5	2.6	2.5	1.6	0.9	0.1	0.3	2.2	3.5	2.0	2.3	1.1	1.4	0.8	1.2	0.6	0.9
Min	0.6	1.2	1.3	1.0	0.7	0.3	0.1	0.1	1.0	1.9	0.9	1.6	0.5	0.7	0.4	0.6	0.3	0.5
Max	2.4	4.8	6.0	5.6	3.5	2.0	0.3	0.5	5.0	8.4	3.6	3.6	1.9	2.4	1.4	1.6	1.1	1.3

[a] Nine of these studies had a NOSTASOT dose lower than the highest experimental dose. Shaded boxes indicate BME options pursued in the entire database.

[b] For change relative to control standard deviation, none of the initial BME options yielded results resembling the NOSTASOT. Therefore, 0.5 standard deviations were selected for application to the entire database, as it was estimated to better approximate the NOSTASOT than either 0.1 or 1 SD.

[c] Values are the ratios of the BMDs to NOSTASOT dose (see Faustman *et al.*, 1994).

Spearman rank correlations were used to explore the similarity among both the various BMDs estimates and the BMD/NOSTASOT ratios (Table 4). For the BMDs, very high correlations ($r > 0.98$) were found among the four litter-based approaches; the correlation between the two quantal-based approaches was even higher ($r > 0.99$). Correlations nearly as strong ($r > 0.95$) were also found between the litter-based and the fetus-based approaches. Correlations between the various BMD/NOSTASOT ratios were generally lower than between the BMDs themselves ($r = 0.58$ to 0.90), probably due to the impact of variable study design (primarily dose spacing) on determination of the NOSTASOT.

This approach to "quantalizing" the continuous data yielded BMDs very similar to the litter-based approaches, although there was evidence that the confidence limits on the maximum likelihood estimates (MLE) were larger for the former approach. Thus, the two fetus-based approaches had average MLE/BMD ratios of 1.96 and 1.83 for the BMEs based upon a 5% added risk of weighing less than the 5th percentile or a 10% added risk of weighing less than 10th percentile, respectively; the average litter-based MLE/BMDs were 1.58 for a 5% decrease in mean weight, 1.62 for a decrease to the 25th percentile, 1.80 for a decrease of 2 SEMs, and 1.71 for a decrease of 0.5 SDs. The averages of the four litter-based MLE/BMDs were significantly different from the average of the two fetus-based approaches (there were significant differences among the litter-based approaches as well).

DISCUSSION

Most attention to date on the application of benchmark dose techniques to toxicology data has focused on quantal measures of response. Before the advantages of the BMD approach can be exploited in the risk assessment process, it is important that continuous measures of response can also be modeled appropriately. In this study, we examined a variety of approaches to estimating BMDs for fetal weight changes as a result of chemical treatment from a total of 85 developmental toxicity experiments. Fetal weight changes are often a very sensitive measure of effect in developmental toxicity studies, thus making it an important endpoint in the risk assessment process. The datasets used here were a subset of those previously used to explore the ability of the BMD approach to model quantal endpoints of response, such as the incidence of prenatal death and/or malformations (Faustman *et al.*, 1994; Allen *et al.*, 1994a,b). Together, the studies demonstrate that the BMD approach can be applied readily to dose–response information obtained from standard developmental toxicity bioassays as used for regulatory purposes and provide a basis for comparing quantitative dose–response estimates with traditionally derived NOAELs.

To apply the BMD approach to a continuous variable such as fetal weight, a definition of what constitutes an affected litter or fetus in terms of a weight decrement had to be developed (i.e., defining the BME). Lacking a clear definition of a biologically significant decrement in fetal weight in a developmental toxicity study, we explored a variety of options based on changes from the control mean fetal weight, as well as on defining individual fetuses, rather than the mean litter response, as being reduced in weight. The location of the percentiles on the cumulative distribution curve that best matched the NOAEL (Fig. 2) tended to be further from the lower inflection for the litter-based BMEs (i.e., the 25th percentile) than for fetus-based BMEs (i.e., the 5th and 10th percentiles). While we were able to develop BMEs for the

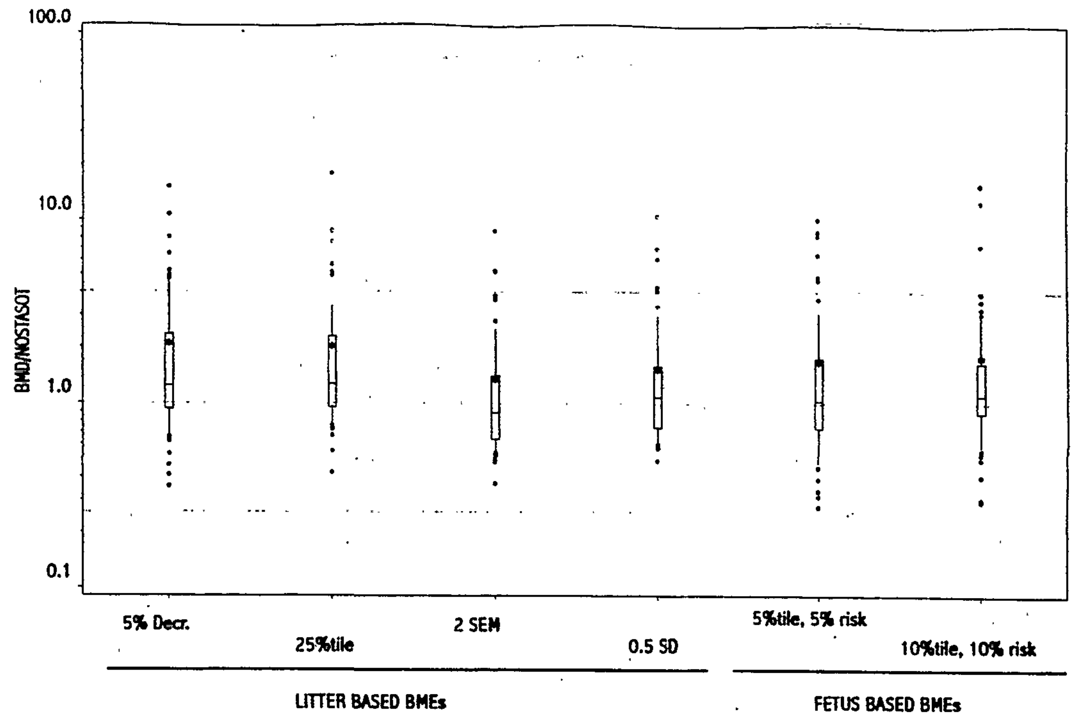

FIG. 3. Distribution of BMD/NOSTASOT ratios for six definitions of a BME. In this box and whisker plot, the boxes contain the interquartile range, the whiskers extend upward to the 90th percentile and downward to the 10th percentiles. Data points outside that range are indicated by an "o". The horizontal line in a box represents the median value, and the "*" in the box represents the mean. Note the influence of the skewness in the ratios, as the means are always higher than the medians. Note also the logarithmic scale of the ordinate. For the *x*-axis label, 5% Decr is based on a 5% reduction in mean litter weight; 25%tile is based on a reduction of the mean litter weight to the 25th percentile of the control distribution; 2 SEM is based on reduction of the mean litter weight by two standard errors of the mean; 0.5 SD is based on one-half a standard deviation decrease in mean litter weight; and 5%tile, 5% risk and 10%tile, 10% risk are the fetus-based benchmarks for a 5% added risk of weighing less than the 5th percentile of the control fetuses, and a 10% added risk of weighing less than the 10th percentile of control fetuses, respectively. There were 76 datasets evaluated for the first four options, 74 for the BME = 5%tile, 5% risk, and 73 for the BME = 10%tile, 10% risk.

litter- and fetus-based BMEs that, on average, gave results comparable with the NOAELS, the confidence intervals about the BMEs for the fetus-based approaches were significantly larger than those for the litter-based approaches. Although the final options all provided BMDs that were similar to each other and had similar ratios to the NOAELs, we recommend against further consideration of the 2 SEM approach because it is influenced by sample size in a manner inconsistent with the primary advantage of the BMD concept in rewarding better experimental design.

Since the six BME options for calculating benchmark doses were selected based on their ability to yield BMDs similar to statistically determined NOAELs, it is perhaps not surprising that they are all highly correlated with one another, as well as with the NOAEL. However, it is interesting to note that the litter-based and fetus-based BMDs consistently yielded results of similar magnitude (i.e., a 5% decrease in mean litter weight yields a dose value that is similar to a dose increasing the proportion of very low weight fetuses by 5%). This suggests that treatments in general are shifting the entire distribution of fetal weights, rather than merely shifting the weights of low weight fetuses to even lower levels. Therefore, while there are theoretical advantages to including both a litter-based and fetus-based BMD in the dose–response assessment due to the possibility of a treatment only affecting the tail of the distribution, such a consideration was not supported by this database.

Given that the vast majority of values for BMDs and NOSTASOTs were within a factor of 2 of each other (the range for the 6 benchmark approaches was 67 to 85%, Figs. 4A–4F), it is worthwhile to explore in greater depth the reasons for those ratios that exhibited larger dissimilarities. Across the six methods of estimating the BMD, a total of nine studies yielded at least one BMD that deviated from the NOSTASOT by fourfold or greater (Table 3). A closer examination of the BMDs from those studies shows that the magnitude of the differences between the BMD and NOSTASOT were relatively consistent within studies. For example, in five of the nine studies, at least four of the six benchmarks differed from the NOSTASOT by more than

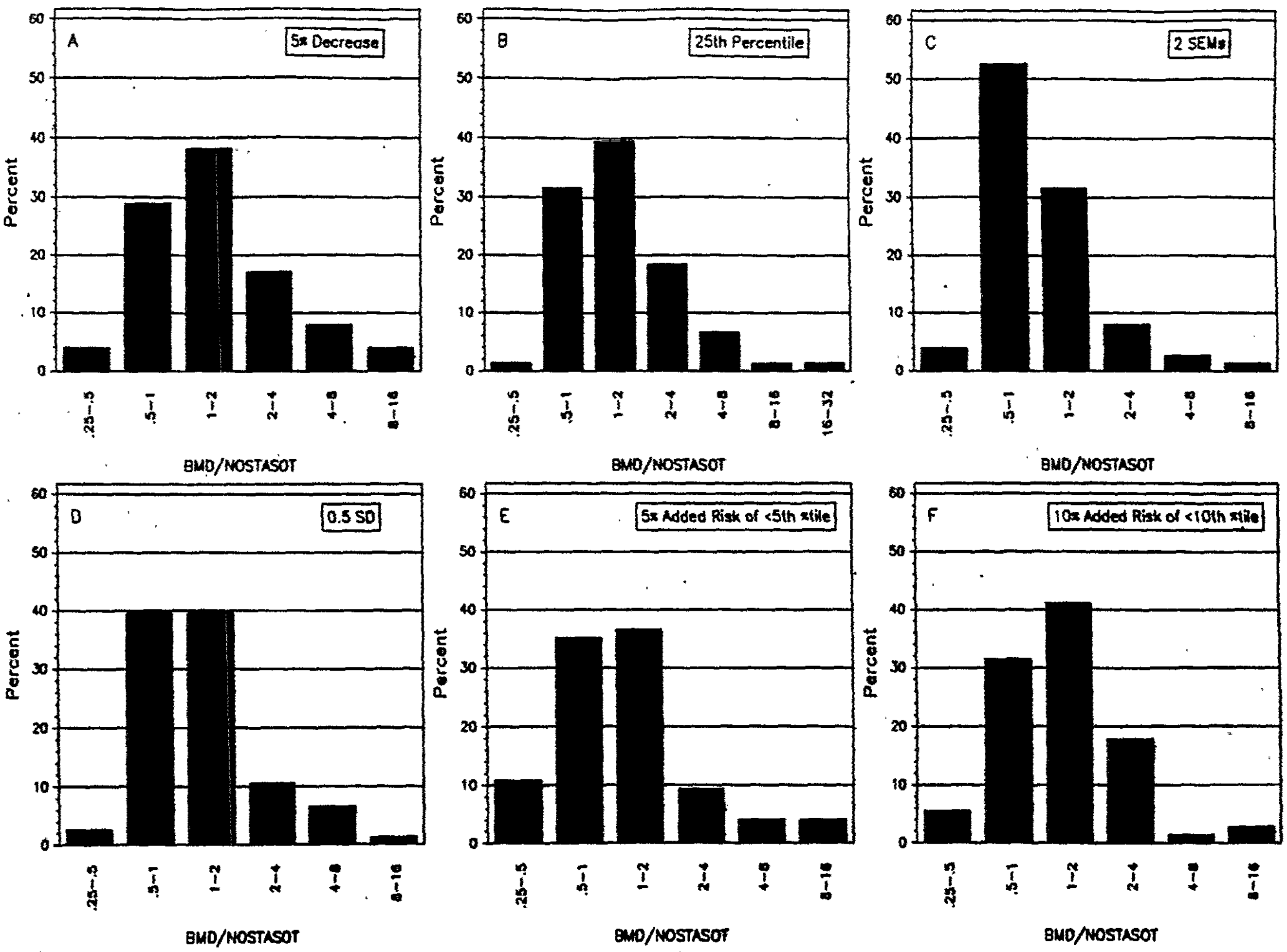

FIG. 4. Frequency histograms of the ratios of the BMD to NOSTASOT for each of the four litter-based and two fetus-based approaches to estimating benchmark doses. Ratios are grouped in categories of a factor of 2. Note that for all approaches, the majority of datasets yield ratios that lie between 0.5 and 2. The distributions tend to be skewed to the right, indicating that when BMDs differed appreciably from the NOSTASOT, they tended to be numerically greater (see Fig. 3 and Results). Note that there were no examples in which the BMD was more than a factor of 4 lower than the NOSTASOT.

fourfold. For one approach (a 5% decrease in mean litter weight), the fourfold difference was observed in all nine datasets. In only one dataset (Study 110 in Table 3) did only one BMD differ from the NOSTASOT by the fourfold factor (the other five BMDs ranged from 1.8- to 3.4-fold different). The average BMD to NOSTASOT value for all six BMD approaches for all nine studies was 5.3 (range 3.0 to 12.3). Thus, whatever contributes to the dissimilarities appears to be more a function of the studies themselves, rather than specific features of the various BME approaches.

Two principal explanations can be offered for these wide dissimilarities: (1) many of these studies had relatively shallow dose–response patterns for fetal weight, making the statistical determination of the NOSTASOT less reliable; and (2) perhaps more importantly, a wide spacing in the dose levels was generally employed. It is common practice in developmental toxicity screening studies to employ an approximate even log spacing (doubling) of the treatment groups. For example, in the database used for our analysis of the BMD approach (Faustman *et al.*, 1994), 189 of the 246 studies (77%) had average spacing between dose groups of between 1.5- and 4-fold; only 46/246 (19%) had average spacing exceeding 4-fold. In contrast, in these nine studies, the average spread between the NOAEL dose and the next highest dose was 9.3. To illustrate, an NTP study (Study 185 in Table 3) in which mice were treated with scopolamine at dose levels of 0, 10, 100, 450, and 900 mg/kg/day, had corresponding mean litter weights of 1.05, 0.97, 1.00, 0.94, and 0.94 g, respectively. The NOSTASOT for this study was the 100 mg/kg/day dose level and the BMD for a reduction in mean litter weight of 5% was 434 mg/kg/day, making the ratio of the BMD to the NOAEL 4.3. Likewise, Study 32

TABLE 3
Studies for Which the BMD Values for Fetal Weight Deviated from the NOSTASOT Dose by at Least a Factor of 4

Study	Litter-based approaches: 5% Decrease	25th %tile	2 SEMs	0.5 SDs	Fetus-based approaches: BMD5–5	BMD10–10	Average
10	8.1[a]	4.9	3.8	4.3	8.3	—[b]	5.9
26	10.7	8.8	5.3	6.2	10.2	15.5	9.5
32	15.1	18.0	8.8	10.6	8.7	12.5	12.3
39	4.6	3.1	2.5	2.8	3.1	4.0	3.4
81	5.3	5.2	3.7	4.1	6.5	7.3	5.4
98	6.5	5.7	3.8	4.3	3.7	3.0	4.5
104	4.7	5.0	2.8	3.4	4.9	3.3	4.0
110	5.0	3.4	2.6	3.0	1.8	2.3	3.0
185	4.3	7.6	5.3	7.0	4.7	3.1	5.3

[a] Values are BMD/NOSTASOT ratios. Cells with shaded boxes contain BMDs which differed from NOSTASOTs by fourfold or greater.
[b] Log–logistic model failed to converge on BME for this dataset.

(Fig. 5) utilized a 10-fold dose spacing between the lowest dose and the middle dose. This study design, combined with a shallow dose–response slope for fetal weight, resulted in various BMD estimates that were much closer to the LOAEL (the middle dose in this study) dose than to the NOSTASOT.

The average effect size (percentage reduction in mean litter weight) at the LOAEL for the nine studies was 4.6% (a relatively small decrement in fetal weight), and for these studies the average BMD tended to be much nearer the LOAEL (BMD/LOAEL ratio 1.1) than the NOAEL (BMD/NOAEL ratio 5.7). Thus, when study designs were used that placed a premium on including very low doses to maximize the possibility of identifying a NOAEL, the NOSTASOT approach seriously overestimated, on average, the potency of the chemical to induce reductions in fetal body weights.

The value of the benchmark dose in instances where no NOAEL could be identified in a study is clearly indicated by our analysis. In nine studies, a significant dose-related

TABLE 4
Spearman Rank Correlations Coefficients between Various Approaches to Deriving Benchmark Dose Levels for Fetal Weight (Correlations between BMDs Shown Shaded; Correlations between BMD/NOSTASOT Ratios Shown Open)[a]

	Litter-based approaches: 5% Decrease	25th percentile	2 SEMs	0.5 SD	Fetus-based approaches: 5th percentile, 5% risk	10th percentile, 10% risk
Litter-based approaches						
5% Decrease	—	0.83	0.81	0.84	0.58	0.63
		76	76	76	74	73
25th percentile	0.98	—	0.85	0.89	0.61	0.70
	85		76	76	74	73
2 SEMs	0.98	0.99	—	0.98	0.63	0.70
	85	85		76	74	73
0.5 SD	0.98	0.99	0.99	—	0.64	0.73
	85	85	85		74	73
Fetus-based approaches						
5th percentile, 5% risk	0.95	0.96	0.97	0.97	—	0.90
	83	83	83	83		71
10th percentile, 10% risk	0.95	0.96	0.98	0.97	0.99	—
	82	82	82	82	80	

[a] The number below the correlation coefficient is the number of comparisons. Note that more comparisons were possible between the BMDs themselves because NOSTASOTs could not be determined for nine studies (see text for details); fewer comparisons were possible between the litter-based approaches than between the litter-based and fetus-based approaches due to the inability of the log–logistic model fit the data for some experiments. All correlations are significant at $p < 0.001$.

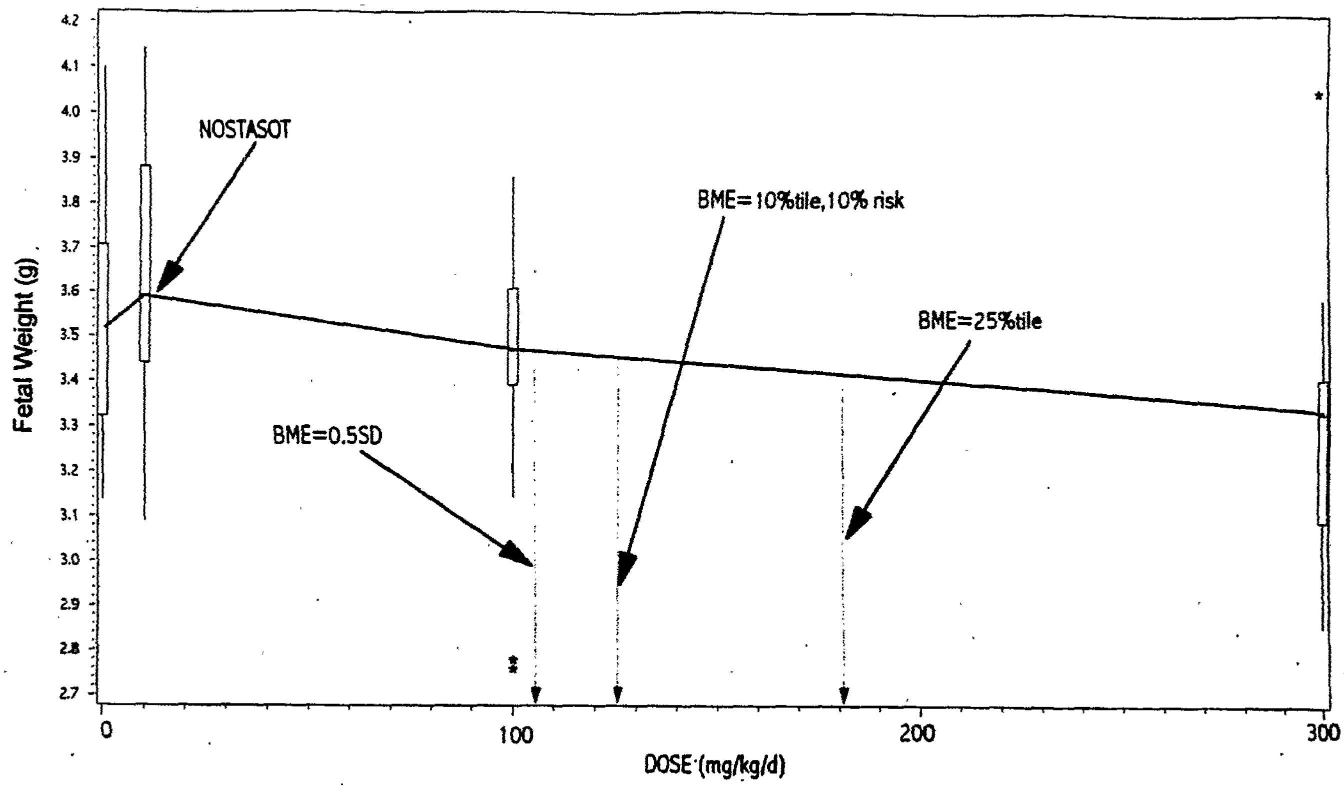

FIG. 5. Example of the impact of study design on the BMD/NOSTASOT ratio. The dose–response for fetal weight from Study 32 (see Table 3) is plotted, and the doses associated with the NOSTASOT, and the highest, lowest, and middle range estimates of the BMDs are indicated by arrows (see legend to Fig. 3 for interpretation of BME designations). In this box and whisker plot, the boxes cover the interquartile range of the mean litter weights, the whiskers cover 1.5 interquartiles in either direction, and outliers are designated by an asterisk. The dose–response curve is drawn through the median values. Note the wide dose spacing (10-fold) between the lowest dose and the middle dose, and the tendency for the BMDs to be near or above the LOAEL dose.

effect on fetal weight was evident but no NOAEL could be established, as the lowest experimental dose produced a significant effect. The default assumption in current risk assessment practice for these studies would be to apply an uncertainty factor of 10 to the lowest experimental dose to estimate the NOAEL. Another option would be to repeat the experiments using lower dose levels. A strength of the benchmark approach is its independence from the confines of the experimental dose groups and its ability to estimate effect levels outside the experimental range. In these nine studies, the average reduction in mean litter weight at the lowest experimental dose (hence defined as the LOAEL) ranged from 0.9 to 11.0% (mean 6.3%), and the range of the BMD/LOAEL for a 5% reduction in mean litter weight was 0.32 to 3.20 (mean 1.53). In three of these studies the BMD for a 5% reduction in mean litter weight was less than the LOAEL (ratios to the LOAEL were 0.66, 0.62, and 0.32). Thus, the BMDs tended to be near or within the experimental dose range, and no large extrapolations of the data were necessary. An example of the estimation of the BMD in a study lacking a NOSTASOT is presented in Fig. 6. Successful estimation of a BMD for these studies precludes the need for application of an additional uncertainty factor or a repeat of the study using lower dose levels.

The process of adopting more quantitative approaches for assessing developmental risks is likely to be iterative in nature. As we acquire information and gain confidence in the more quantitative, new considerations are sure to arise that may warrant additional research. From our viewpoint, such additional work is suggested in several areas. These include: (1) The overall impact of the BMD approach on determination of the critical effect in a given study. To date, we have only compared endpoint-specific NOAELs to corresponding BMDs, and have not taken the next step of comparing study-based NOAELs with study-based BMDs. (2) The application of multivariate models that consider both quantal (incidence) and continuous (weight) endpoints simultaneously, such as proposed by Catalano *et al.*, (1993), to the same database used here to determine the differences in that approach to BMD estimation over the present univariate approaches. (3) Evaluation of whether current study designs that were optimized for determination of NOAELs can be improved upon

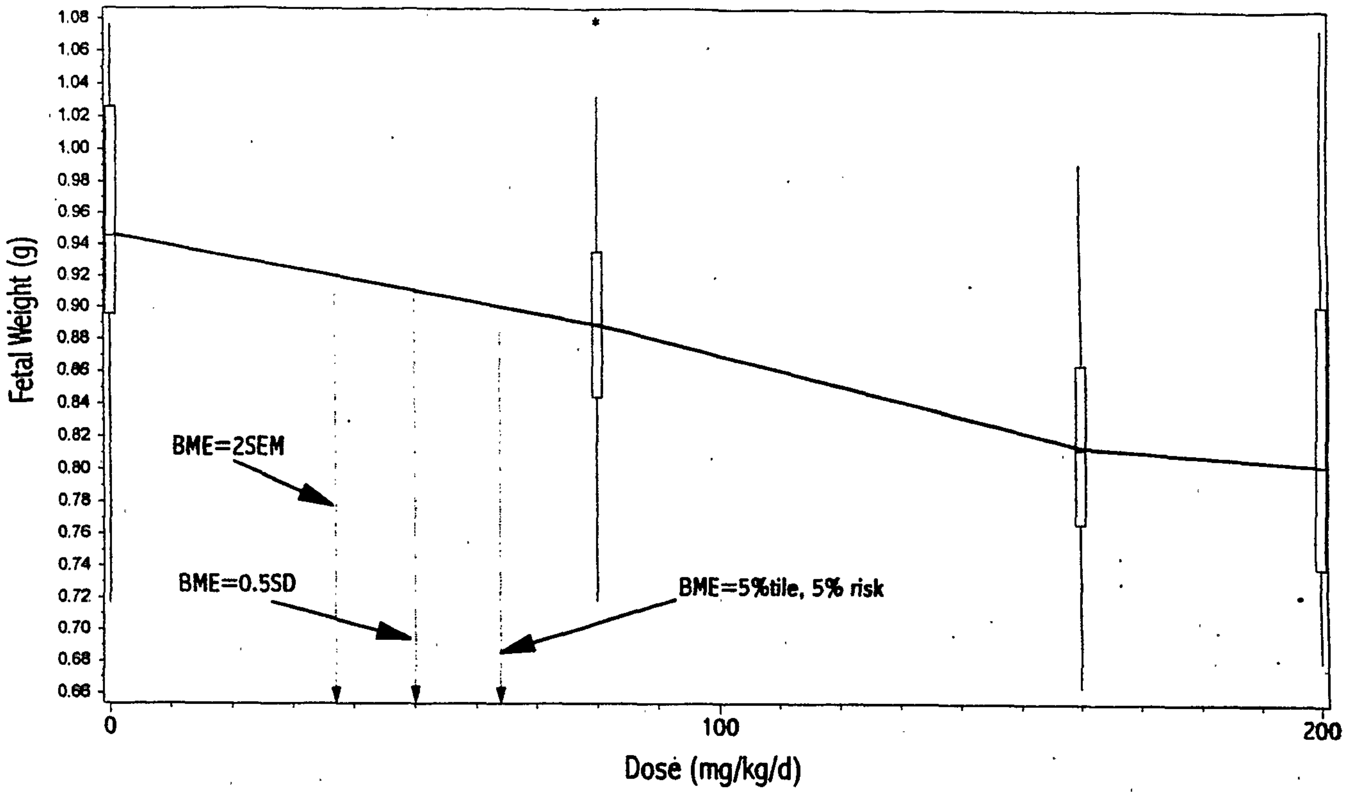

FIG. 6. Illustration of BMD estimation from study with no NOSTASOT. The dose–response data for fetal weight from Study 181 is plotted, and the doses associated with the highest, lowest, and middle range estimates of the BMDs are indicated by arrows (see legend to Fig. 3 for interpretation of BME designations). In this box and whisker plot, the boxes cover the interquartile range of the mean litter weights, the whiskers cover 1.5 interquartiles in either direction, and outliers are designated by an asterisk. The dose–response curve is drawn through the median values. Note that the BMDs cluster between the control dose level and the lowest dose level.

for estimating BMDs. Efforts in this area are underway (Kavlock and Schmid, 1994; Weller *et al.*, 1994). And finally, (4) the development of mechanistically based dose–response models.

Combined with our previous publications (Faustmen *et al.*, 1994; Allen *et al.*, 1994a,b), we have now examined the application of the benchmark dose approach for all fetal endpoints of developmental toxicity and have provided a firm basis for its use in the risk assessment process. The introduction of statistical models for providing quantitative risk estimates is only an initial step in the ultimate goal of developing biologically based quantitative risk estimations that incorporate information on species-specific pharmacokinetics, mechanistic interactions, and pathogenesis. Research in this area is only in its infancy (Gaylor and Razzaghi, 1992; Shuey *et al.*, 1994; Leisenring *et al.*, 1994; Setzer, 1994).

ACKNOWLEDGMENTS

We thank Dr. Bern Schwetz (NCTR) for the contribution of the NTP datasets and Dr. Joseph Holson, Mark Nemec, and Stan Kopp for the contributions of the WIL Laboratories datasets. Judy Schmid of ManTech Environmental provided assistance with the descriptive statistics and graphical representations. We also appreciate the thoughtful suggestions of Drs. Woody Setzer and John Vandenberg of the U.S. EPA. This research was supported by a cooperative agreement between the U.S. Environmental Protection Agency and the University of Washington.

REFERENCES

Allen, B. C., Kavlock, R. J., Kimmel, C. A., and Faustman, E. F. (1994a). Dose–response assessments for developmental toxicity. II. Comparison of generic benchmark dose estimates with NOAELs. *Fundam. Appl. Toxicol.* **23**, 487–495.

Allen, B. C., Kavlock, R. J., Kimmel, C. A., and Faustman, E. F. (1994b). Dose–response assessments for developmental toxicity. III. Statistical models. *Fundam. Appl. Toxicol.* **23**, 496–509.

Barnes, D. G., Daston, G. P., Evans, J. S., Jarabek, A. M., Kavlock, R. J., Kimmel, C. A., Park, C., and Spitzer, H. L. (1995). Benchmark dose workshop: Criteria for use of a benchmark dose to estimate a reference dose. *Regul. Toxicol. Pharmacol.* **21**, 296–306.

California EPA (1994). *Safety Assessment for Non-Cancer Endpoints: The Benchmark Dose and Other Possible Approaches.* Summary report on a workshop held May 11–12, 1992, Reproductive and Cancer Hazard Section, Office of Environmental Health Hazard Assessment, June 1994.

Catalano, P. J., Scharfstein, D. O., Ryan, L. M., Kimmel, C. A., and Kimmel, G. L. (1993). Statistical model for fetal death, fetal weight and malformation in developmental toxicity studies. *Teratology* **47**, 281–290.

Cox, D. R., and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, New York.

Crump, K. S. (1984). A new method for determining allowable dietary intakes. *Fundam. Appl. Toxicol.* **4**, 854–871.

Crump, K. S., and Howe, R. (1985). A review of methods for calculating confidence limits in low dose extrapolation. In *Toxicological Risk Assessment* (D. Krewski, Ed.). CRC Press, Canada.

Faustman, E. F., Allen, B. C., Kavlock, R. J., and Kimmel, C. A. (1994). Dose–response assessment for developmental toxicity. I. Characterization of database and determination of NOAELs. *Fundam. Appl. Toxicol.* **23**, 478–495.

Gaylor, D. W. (1989). Quantitative Risk Analysis for Quantal Reproductive and Developmental Effects. *Environmental Health. Perspectives* **79**, 243–246.

Gaylor, D. W., and Razzaghi, M. (1992). Process of building biologically based dose–response models for developmental defects. *Teratology* **46**, 573–581.

Gaylor, D. W., and Slikker, W., Jr. (1990). Risk assessment for neurotoxic effects. *Neurotoxicology* **11**, 211–218.

Haseman, J. K. (1984). Statistical issues in the design, analysis and interpretation of animal carcinogenicity studies. *Environ. Health. Perspect.* **58**, 385–392.

Kavlock, R. J., and Schmid, J. (1994). Influence of study design on benchmark dose calculations. *Teratology* **49**(5), 394.

Kimmel, C. A. (1990). Quantitative approaches to human risk assessment for noncancer health effects. *Neurotoxicology* **11**, 189–198.

Kimmel, C. A., and Gaylor, D. W. (1988). Issues in qualitative and quantitative risk analysis for developmental toxicity. *Risk Anal.* **8**, 15–20.

Kupper, L., Portier, C., Hogan, M., and Yamamoto, E. (1986). The impact of litter effects on dose–response modeling in teratology. *Biometrics* **42**, 85–98.

Leisenring, W. M., Leroux, B. G., Moolgavkar, S. H., Ponce, R. A., and Faustman, E. M. (1994). A biologically based dose–response model for the developmental toxicity of methyl mercury. *Toxicologist* **14**(1), 39.

Pease, W., Vandenberg, J., and Hooper, K. (1991). Comparing alternative approaches to establishing regulatory levels for reproductive toxicants: DBCP as a case study. *Environ. Health Perspect.* **91**, 141–155.

Setzer, R. W. (1994). Development of biologically-based dose response models: Using a méchanistic model to generate experimentally testable hypotheses about mechanisms of the developmental toxicity of 5-FU. *Teratology* **49**(5), 396–397.

Shuey, D., Lau, C., Logsdon, T. R., Zucker, R. M., Elstein, K. H., Narotsky, M. G., Setzer, R. W., Kavlock, R. J., and Rogers, J. M. (1994). Biologically based dose–response modeling in developmental toxicology: Biochemical and cellular sequelae of 5-fluorouracil in the developing rat. *Toxicol. Appl. Pharmacol.* **126**, 129–144.

Tukey, J., Ciminera, J., and Heyse, B. (1985). Testing the statistical certainty of a response to increasing dose of a drug. *Biometrics* **41**, 295–301.

Weller, E. A., Catalano, P. J., and Ryan, L. M. (1994). Implications of various animal allocation and dose spacing designs for quantitative risk assessment. *Teratology* **49**(5), 366.